Fundamentals of Big Data Application

基　础　篇

BIG DATA

第 1 章 大数据概述

1.1 数据和大数据

1.1.1 数据的高速增长

数据（data）是可以定量分析的记录。远古时代，人们“结绳记事”来记载事件，后来为了记录更复杂的事件（如白天时长、气候变化等），创造了某种抽象的符号，最后逐渐发展成为不同类型的数字。中国是最早进行国情调查的国家之一。《史记》曾记载“禹平水土，定九州，计民数”，可以判断大禹平定水患之后，就在九州范围内统计人口了。以后的朝代也有较大规模的人口统计，如明洪武年间编制的黄册（全国户口名册）。统计学起源于国情调查，是搜集、整理、分析、解释数据并从数据中得出结论的一门科学。在统计学理论的指导下，人们通过市场调查、产品抽查和控制实验等方式，对总体进行抽样并获得数据，基于概率理论、抽样误差等理论建立数学模型，估计和检验总体的参数并用于预测。

20 世纪末，信息技术在社会各个领域都得到了全面应用，数据的产生、存储、传输和利用方式都产生了巨大变化，人类对数据的处理水平也空前提高。现在，数据不再仅指数值，还包括文字、图像、图形、动画、视频、音频等多种表现形式。21 世纪以来，互联网全面融入了经济社会生产和生活的各个领域，引领了生产方式和生活方式的变革，创造了人类生活新空间，并深刻地改变着全球产业、经济、利益、安全等的格局。移动互联网、智能终端、新型传感器快速渗透到生产和生活中，一个与物理世界平行的数字空间正在形成，数据成为继物质、能源之后的又一种重要战略资源。

互联网，特别是移动互联网的快速普及，使个人成为重要的数据生产者。当一个人发送消息或图片、电子邮件或搜索信息或提交报表时，就已经开始了数据生产。2021 年，全世界 46.68 亿互联网用户每天发起了 50 亿次在线搜索，发送了 3000 亿封电子邮件，平均每人产生了 4000 次数据互动（主动和被动）。物联网技术在工业领域得到了深入应用，数以亿计的传感器、边缘计算设备和智能设备夜以继日地产生着

生产数据。

随着全球数据种类的不断增多，数据总量也在以令人难以置信的速度持续增长。为了计量数据的大小，人们创造了表 1.1 所示的计量单位。其中，最小的数据计量单位是比特或位 (b)，8 个比特形成一个字节 (B)，1024 个字节形成一个千字节（KB），1024 个千字节形成一个兆字节（MB）。以 1024 为倍数，依次形成吉字节（GB）、太字节（TB）、拍字节（PB）、艾字节（EB）、泽字节（ZB）和尧字节（YB）。事实上，TB 及以上的单位能够计量的数据，已经超出了非专业人士的想象范围。

表 1.1　数据计量单位

存储单位	换算关系	含义与实例
b（比特或位）		1 b 是指一个二进制数 (1 或 0)
B（字节）	1 B=8 b	可以表达西文字符
KB（千字节）	1 KB=1024 B=2^{10} B	1 页纯文字的 Word 文件的大小均为 15 KB
MB（兆字节）	1 MB=1024 KB=2^{20} B	一个 MP3 格式的音乐文件的大小均为 4 MB
GB（吉字节）	1 GB=1024 MB=2^{30} B	一部高清电影的大小约为 1 GB
TB（太字节）	1 TB=1024 GB=2^{40} B	2022 年，主流笔记本电脑的硬盘容量为 1 TB
PB（拍字节）	1 PB=1024 TB=2^{50} B	人类生产的所有印刷材料的数据量为 200 PB
EB（艾字节）	1 EB=1024 PB=2^{60} B	人类说过的语言的总和的大小均为 5 EB
ZB（泽字节）	1 ZB=1024 EB=2^{70} B	2025 年，全球数据量预计达到 500 ZB
YB（尧字节）	1 YB=1024 ZB=2^{80} B	2029 年，全球数据量预计达到 1 YB

2005 年，全球产生的数据量为 130 EB；2010 年的数据量为 1 ZB；2015 年，全球数据量达到近 15 ZB；2020 年达到 50 ZB。迄今为止，人类生产的所有印刷材料的数据量为 200 PB，全人类历史上说过的所有语言的数据量大约为 5 EB。整个人类文明所获得的全部数据中，有 90% 是过去两年内产生的，数据呈现出以几何级数增长的趋势。

我国有超过 10 亿人在使用互联网，中国已成为全球数据总量最大、数据类型最丰富的国家之一。

1.1.2　大数据

大多数学者认为，“大数据”这一概念最早公开出现于 1998 年。美国高性能计算公司 SGI 的首席科学家约翰 • 马西（John Mashey）在一个国际会议报告中指出：随着数据量的快速增长，必将出现数据难理解、难获取、难处理和难组织等四个难题，并用“big data（大数据）”来描述这一挑战，在计算领域引发思考。2008 年 9 月，《自然》杂志出版了以“大数据”为主题的专刊。2011 年 2 月，《科学》杂志出版了专刊——*Dealing with data*，开篇文章发布了一个关于数据使用的调查结果：91.2% 的人认为无法有效驾驭所拥有的数据。世界著名的管理咨询公司麦肯锡公司于 2011 年 5

月发布了一份题为《大数据：竞争、创新和生产力的下一个前沿》的报告。该报告认为，所谓大数据是指“规模已经超出典型数据库软件所能获取、存储、管理和分析能力之外的数据集。”报告提出了对大数据进行收集和分析的设想，并对大数据会产生的影响、所需关键技术以及应用领域等进行了较详尽的分析。2012 年，牛津大学教授维克托·迈尔 - 舍恩伯格（Viktor Mayer-Schönberger）在其畅销著作《大数据时代：生活、工作、思维的大变革》（*Big Data: A Revolution That Will Transform How We Live, Work, and Think*）中指出，数据分析将从“随机采样”“精确求解”和“强调因果”的传统模式演变为大数据时代的“全体数据”“近似求解”和“只看关联不问因果”的新模式，引发了商业应用领域对大数据方法的广泛思考与探讨。

国际数据公司（International Data Corporation，IDC）认为大数据具有海量的数据规模、快速的数据流转、多样的数据类型和较低的价值密度四大特征。

1. 海量的数据规模

大数据集往往能够达到 TB 甚至 PB 数量级。例如，导航软件每天需要处理的数据超过 1.5 PB。由于数据体量巨大，传统的存储技术和处理技术不再适用。例如，传统的数据处理方法对京东、天猫等电商网站一天产生的交易数据是无能为力的。

2. 快速的数据流转

数据生成的速度非常快。例如，工地或车间的摄像头会高速地产生大量数据，大型强子对撞机（Large Hadron Collider，LHC）在工作状态下每秒产生 PB 级的数据。

3. 多样的数据类型

大数据的来源和类型是多种多样的。例如，一个企业的数据可能包括财务和生产的数值数据，也包括电子邮件、文档、社交媒体、图片、音频和视频等数据。而交通领域的数据则可能包含路网摄像头、传感器、GIS（geographic information system，地理信息系统）数据、问卷调查、交通卡刷卡记录、手机定位记录、高速公路及停车场 ETC（electronic toll collection，电子不停车收费）数据等不同来源和类型的数据。

4. 较低的价值密度

以视频监控为例，监控数据量非常大，但有用的数据可能只有几秒。不过，数据量大并不是导致其价值密度低的原因，银行、大型电子商务网站的海量交易数据和医院病历数据都有较高的价值。

关于大数据，研究机构 Gartner 给出了这样的定义：大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合。必须使用高级工具（分析和算法）进行处理，才能从大数据中揭示有意义的信息。

经过多年来的发展和沉淀，人们对大数据已经形成了基本共识：大数据现象源于

互联网及其延伸所带来的无处不在的信息技术应用以及信息技术的不断低成本化。大数据泛指无法在可容忍的时间内用传统信息技术和软硬件工具对其进行获取、管理和处理的巨量数据集合，具有海量性、多样性、时效性及可变性等特征，需要可伸缩的计算体系结构以支持其存储、处理和分析。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有价值的数据进行专业化处理。大数据的价值本质上体现为：为人类提供了全新的思维方式和探知客观规律、改造自然和社会的新手段，这也是大数据引发经济社会变革的根本性原因。

1.1.3 科学的范式

1962 年，美国著名科学哲学家托马斯·塞缪尔·库恩（Thomas Samuel Kuhn）在《科学革命的结构》中提出了“范式”（paradigm）这一概念。范式指的是常规科学所赖以运作的理论基础和实践规范，是从事某一学科的科学家群体所共同遵从的世界观和行为方式。新范式的产生，一方面是由于科学范式本身的发展；另一方面则是由于外部环境的推动。人类进入 21 世纪以来，随着信息技术的飞速发展，新的问题不断产生，原有的科学范式受到了各个方面的挑战。

在科学发展史上，第一范式是实验科学（经验科学）。实验科学的基本特征是对有限的客观现象进行观察、总结、提炼，用归纳法发现科学规律。伽利略利用实验和数学相结合的方法确定了一些重要的力学定律。他对落体运动进行了细致观察之后，在比萨斜塔上做了“两个铁球同时落地”的著名实验，得出了“物体下落速度与重量无关”的结论。实验科学的主要研究模型是科学实验，其研究方法以归纳为主，观测和实验带有一定的盲目性。

第二范式指 18 世纪以来的理论科学。理论科学的主要活动是对自然、社会现象按照已有的实证知识、经验、事实、法则、认知以及经过验证的假说，经由一般化与演绎推理等方法，进行合乎逻辑的推论性总结。例如，相对论、麦克斯韦方程组、量子力学、概率论、博弈论等均属于理论科学范畴。理论科学的主要模型是数学模型，其研究方法以演绎法为主，不局限于经验事实。

第三范式指 20 世纪中叶以来的计算科学。面对大量复杂现象，归纳法和演绎法都难以满足科学需求，人们开始借助电子计算机对复杂现象进行模拟仿真，并推演出越来越多复杂的现象，以完成观察和预测。例如，分子问题、信号系统均属于计算科学的范畴。计算科学的主要研究模型是计算机仿真和模拟，其研究方法是针对问题进行仿真计算。

近年来，人类拥有的数据以惊人的速度增长，传统的计算科学范式已经越来越无法驾驭海量数据。图灵奖得主、关系型数据库、数据仓库和数据挖掘方向的领军人物詹姆斯·尼古拉·格雷（James Nicholas Gray）于 2007 年 1 月 11 日在他人生中最后

一次[①]演讲《e-Science：一种科研模式的变革》中指出，科学的发展正在进入数据密集型科研——科学史上的“第四范式”。第四范式开始用超大规模（mega-scale）和细微规模（mili-scale）等概念来描述数据时代的特征，已经具备了大数据的核心内涵。他认为，数据爆炸对传统研究工具提出了挑战和颠覆，需要变革研究工具才能有效利用海量数据。第四范式和大数据革命的思想非常吻合，可以视为大数据革命的思想雏形。

微软研究院于 2009 年出版了《第四范式：数据密集型科学发现》（*The Fourth Paradigm: Data-intensive Scientific Discovery*），标志着第四范式——数据密集型科学范式的确立。数据密集型科学的研究对象是全量数据，而不是少量样本。计算机科学为信息分析学科等提供了研究的手段和工具，已经成为众多学科必需的辅助科学。数据密集型科学的研究不再由传统的假设驱动，而是基于科学数据进行探索，主要研究方法为数据挖掘。

1.2 大数据从哪里来

传统数据是按照特定研究目的、依据抽样方法获得的格式化的数据。而大数据的产生主体则有以下几个。

组织的信息系统产生数据。组织内部使用的信息系统，如企业资源规划、制造执行系统、医院信息管理系统、办公自动化系统等，会产生运营数据（如产品数量、原材料消耗等）和系统二次加工（如分类、汇总）数据。

用户产生数据。互联网上，社交媒体每时每刻都产生大量文字、图片、视频，这些数据还可能包含着互联网用户的行为信息。例如，电商平台记录了用户的支付行为、查询行为、购买习惯、单击顺序、停留时间、评价行为、物流信息等数据。此外，搜索引擎、导航、电子邮件、短信、共享服务（如汽车、自行车）、手机拍照（很大一部分传到了云端）也会产生大量的数据。

机器产生数据。智能设备和传感器在持续不断地产生着数据。例如，智能起重机记录着每个动作的工作时间、起吊重量、环境风力甚至驾驶员的个人生理信息等数据，以监控操作者及设备健康状况，提高安全生产水平；传感器持续记录着温度、湿度、粉尘等实时信息；无人机和摄像头也从各个角度拍摄、记录着现场数据。即使这些数据仅有一部分被保留下来，长期累积的数据量也是惊人的。

科学实验产生数据。航空航天、海洋监测、天气观测、电子对撞机等科学实验会产生海量数据。

① 2007 年 1 月 28 日，Gray 独自驾驶着一条长 40 英尺的游艇，驶往位于旧金山金门大桥以西 25 英里的费拉隆（Farallon）岛途中失踪。

1.3 大数据的应用场景

大数据技术产生于互联网领域，并逐步推广到电信、医疗、金融、交通等领域，在众多行业中产生了实用价值。

1. 互联网领域

互联网企业获取大量的客户行为信息，通过大数据技术分析，可以制定出具有针对性的服务策略，从而获取更大的效益。近年来的实践证明，合理地运用大数据技术能够将电子商务的营业效率提高 60% 以上。电商平台会通过大数据技术采集有关客户的各类数据，使用大数据分析技术建立“用户画像”来描述一个用户的信息全貌，从而对用户进行个性化推荐、精准营销和广告投放等。例如，当用户登录电商网站时，系统就能根据“用户画像”预测出该用户今天可能购买的物品，然后从商品库中把合适的商品找出来，用“猜你喜欢”的方式推荐给他；如果顾客的购物车中有多包羊肉片、糖蒜却没有火锅蘸料，则在结账时询问是否需要蘸料，得到肯定答复后，会将顾客引导到蘸料页面。广告是互联网领域常见的盈利方式，也是一个典型的大数据应用。广告系统能够根据用户的历史行为模式及个人基本信息，针对用户投放精准的广告。

2. 电信领域

在电信行业中，用户每天产生的语音、短信、流量和宽带数据是体量巨大的数据资源。通过大数据技术，运营商可以提升数据处理能力，聚合海量数据，提升洞察能力。目前电信领域主要将大数据应用在以下几个方面。

（1）网络管理和优化。通过数据分析，对基站选址等基础设施建设进行优化；对已有设施进行效率和成本评估，以减少浪费。

（2）市场与精准营销。包括客户画像、关系链研究、精准营销、实时营销和个性化推荐，提高营销效率。

（3）客户关系管理。包括客服中心优化和客户生命周期管理。当前，在中国的电信市场中，各运营商的市场份额比较稳定，防止客户流失是一项重要业务。通过数据分析，如发现客户有“离网”倾向，就可以制定有针对性的措施挽留客户。

（4）企业运营管理。基于企业内部的业务和用户数据，以及通过大数据手段采集的外部社交网络数据、技术和市场数据，对业务和市场经营状况进行总结和分析。

（5）数据商业化。通过与第三方合作，将数据价值外部变现，包括数据即服务和分析即服务。数据即服务是通过开放数据或 API（application programming interface，应用程序编程接口），向外出售脱敏后的数据；分析即服务是指与第三方公司合作，利用脱敏后的数据为政府、企业或行业客户提供通用信息、数据建模、数据分析服务。

3. 医疗领域

在传统的医疗诊断中，医生仅可依靠目标患者的信息以及自己的经验和知识储备，局限性很大。医疗行业拥有病历、病理报告、影像数据、治疗方案、药物报告等数据，数据量庞大并且类型复杂。通过机器学习算法可以发现数据中蕴含的规律，协助医疗团队建立疾病模型。生物大数据的应用能够在很大程度上帮助研究人员调查疾病与人体遗传标记之间存在的必然联系，改变传统医疗模式下对所有的病人都采取“一刀切”的治疗方法，将基因学的内容引入到临床治疗中，对患者的基因组数据进行分析，从而提供针对性的医疗方法。这为后续医疗技术的进步以及疾病预防工作的开展提供了有效的技术支持。

重大流行性疾病防控的关键是发现病毒感染者和密切接触者，通过收治和物理隔离手段切断传染链，其基础工作在于开展科学、准确的流行病学调查，掌握流行病病例的发病情况、暴露史、接触史等流行病学相关信息，而这一切离不开大数据的支撑。流行病学现场调查广泛运用互联网、大数据等技术，对当事人或知情者提供的有效信息进行甄别和综合梳理分析，不仅可以准确掌握当事人的数据信息，甚至其准确度可能比当事人本人直接提供的还要高。正如李兰娟院士所说：“专家利用大数据技术梳理感染者的生活轨迹，追踪人群接触史，成功锁定感染源及密切接触人群，为疫情防控提供宝贵信息。”甘肃省利用公安“天眼”系统和大数据平台调取相关数据，根据与基层流行病学调查组的比对，翔实核查出已确诊患者和疑似病例的活动范围及接触人群，缩短了流行病学的调查时间，拓展了排查渠道，提高了调查结果的准确性。

4. 金融领域

银行拥有多年的数据积累，目前已经开始尝试通过大数据来驱动业务运营。银行大数据应用可以分为四方面。

（1）客户画像应用。客户画像应用主要分为个人客户画像和企业客户画像。个人客户画像包括人口统计学特征、消费能力、兴趣、风险偏好等；企业客户画像包括企业的生产、流通、运营、财务、销售、客户、相关产业链上下游等数据。

（2）精准营销。在客户画像的基础上，银行可以有效地开展精准营销。银行可以根据客户的喜好进行服务或者银行产品的个性化推荐，如根据客户的年龄、资产规模、理财偏好等，对客户群进行精准定位，分析出其潜在的金融服务需求，进而有针对性地进行营销推广。

（3）风险管控。风险管控包括中小企业贷款风险评估和欺诈交易识别等。银行可以利用持卡人基本信息、银行卡基本信息、交易历史、客户历史行为模式、正在发生的行为模式（如转账）等，结合智能规则引擎（如从一个不经常出现的国家为一个特有用户转账或在一个不熟悉的位置进行在线交易）进行实时的交易反欺诈分析。

（4）运营优化。运营优化包括市场和渠道分析优化、产品和服务优化等。通过大数据，银行可以监控不同的市场推广渠道，尤其是网络渠道推广的质量，从而进行

合作渠道的调整和优化；银行可以将客户行为转换为信息流，并从中分析客户的个性特征和风险偏好，更深层次地理解客户的习惯，智能化分析和预测客户需求，从而进行产品创新和服务优化。

5. 工业领域

工业大数据应用将带来工业企业创新和变革的新时代。在工业生产中，信息系统、传感器和智能设备时刻产生着海量数据。工业大数据的典型应用包括产品创新、产品故障诊断与预测、工业生产线物联网分析、工业企业供应链优化和产品精准营销等诸多方面。

（1）加速产品创新。客户与工业企业之间的交互和交易行为将产生大量数据。挖掘和分析这些客户动态数据，能够帮助客户直接参与到产品的需求分析和产品设计等创新活动中，为产品创新做出贡献。

（2）产品故障诊断与预测。传统上，设备会定期检修。无论设备状态如何，都要按计划检修，这会造成一定程度上的浪费。此外，设备出现故障后会立即停机，然后进行故障定位和排除，这种非计划停机会严重影响生产活动。通过传感器监控设备多个指标的实时状态，分析发生故障之前的参数变化规律，就可以建立大数据模型，做到故障预警。这样，维修人员就能在设备停机之前对其进行维修处理，从而提高整体的企业运行效率。

（3）基于数据的产品价值挖掘。通过对产品和相关数据进行二次挖掘，可以创建新的价值。例如，三一重工是我国著名的工程机械供应商，该厂可以在线跟踪它出售（或出租）的设备的工作状态。这些大数据可以帮助客户预防故障，帮助本厂的设计部门改善产品设计，还能了解全国各地的基础设施状况，为宏观经济判断、市场销售布局和金融服务提供依据。

另外，大数据在国家安全、社会治理、体育娱乐、交通管理等领域也有着深入的应用。可以预见，未来大数据会在更多的领域得到深入应用，促进社会发展，造福民众。

1.4 大数据对思维方式的影响

维克托·迈尔·舍恩伯格在《大数据时代：生活、工作、思维的大变革》中明确指出，大数据时代最大的转变是三种转变：总体而非抽样、效率而非精确和相关而非因果。

（1）总体而非抽样。过去，由于数据存储和处理能力的限制，在数据分析中，通常采用抽样的方法，即从全集数据中抽取一部分样本数据，通过对样本数据的分析来推断全集数据的总体特征。现在，获取数据的能力空前提高，分布式文件系统和分布式数据库技术提供了理论上近乎无限的数据存储能力，分布式并行程序设计框架 MapReduce 提供了强大的海量数据并行处理能力。因此，在大数据技术时代，科学分析可直接针对全集数据而不是抽样数据，并且可以在短时间内得到分析结果。

（2）效率而非精确。在基于抽样的分析方法中，抽样的微小误差被放大到全集数据以后，可能会变成一个很大的误差。传统的数据分析方法往往更注重提高数据和算法的精确性，其次才是提高算法效率。而在大数据背景下，基于数据总体的分析结果就不存在误差被放大的问题。因此，追求高精确性已经不是其首要目标。大数据有变化快的特征，要求在几秒内就给出针对海量数据的实时分析结果，否则就会丧失数据的价值。因此，数据分析的效率成为关注的核心。

（3）相关而非因果。在大数据时代，因果关系不再那么重要（事实上，因果关系更难以发现），人们转而追求“相关性”。例如，在电商平台购买了一本 Python 相关的图书后，系统会自动提示与你购买相同物品的其他客户还购买了某本关于大数据的图书。尽管两者可能有一定的因果性，但系统只需要根据相关性做出提醒，而不必发现其中的因果关系。

1.5 数据挖掘与机器学习

在科学和工程研究中，“第一性原理”（first principle）是一个重要准则，即研究要从基本的数学定理或物理规律出发来计算和推导，直到得到结论。例如，在自由落体中，距离和时间的关系用 $h=\frac{1}{2}gt^2$ 表示，给定一个 t，就可以预测 h 的值。

但是，如果想知道人 1 小时能走多少米，问题就变得十分复杂。结论和人的身高、体重、肌肉力量、关节健康等有关系，这个问题可能需要很多参数和方程才能得到数学模型。更准确地说，根本无法得到这样的数学模型。但如果测量了 1000 个人 1 小时走过的距离，就可以得到有一定精度的经验公式，从而估计某种身高体重的人 1 小时大约能走多少距离[①]。和依赖数学或物理公式的第一性原理截然不同，这种预测能力来自数据分析。

20 世纪下半叶，随着数据库技术的发展应用，数据的积累不断膨胀，导致简单的查询和统计已经无法满足企业的商业需求，亟需一些革命性的技术去挖掘数据背后的信息。同时，计算机领域的人工智能也取得了巨大进展，机器学习发挥了重要作用。因此，人们将两者结合起来，用数据库管理系统存储数据，用计算机分析数据，尝试挖掘数据背后的信息。这两者的结合促生了一门新的学科，即数据库中的知识发现（knowledge discovery in database，KDD）。后来，数据来源不再局限于数据库，这个术语逐渐被数据挖掘（data mining）替代。

数据挖掘指从大量数据中通过算法和分析工具获得隐藏于其中的信息的过程，即从大量的、不完全的、有噪声的、随机的、模糊的数据中，提取隐含在其中的规律性的、

① 导航软件一般按 5 km/h 估算步行时间。如果多次使用步行导航，软件会“学习”到用户的步行速度，就能更精准地估计步行时间。

人们事先未知但又是潜在的有用信息和知识的过程。数据挖掘可以帮助决策者寻找数据间潜在的某种关联，发现被隐藏的、被忽略的因素。数据挖掘和统计学的目标都是发现数据中的信息，但是数据挖掘的工作对象不是通过抽样获得的样本，而是来自数据库或网络的数据总体。数据挖掘通过计算机执行算法，以数据驱动的方式发现数据中的信息，为决策提供支持。

机器学习（machine learning）理论主要用于设计和分析一些让计算机可以根据现有数据自动“学习”的算法，并据此建立模型。机器学习算法是从数据中自动分析获得规律，与模型构建有关，而数据挖掘与知识发现有关，两者不是并列的概念。机器学习不但广泛应用于数据挖掘，还应用于计算机视觉、自然语言处理、生物特征识别和机器人领域；而数据挖掘除了机器学习，还涉及数据库理论、人工智能和现代统计学。

第一性原理和数据挖掘以两种不同的方式表达了知识，两者并不矛盾，数据挖掘也能找到一些原理和机制。康奈尔大学（Cornell University）的科学家做过一个有趣的研究，他们从实验数据出发，通过数据挖掘，反向得到了与牛顿力学中一致的物理公式，而计算机并不需要知道牛顿力学。将来，随着算法的进步和算力的提高，人们可能从更复杂的数据中挖掘出未知的规律。

和传统数据相比，大数据的体量更大，价值密度更低，蕴含的知识更加模糊。传统的数据挖掘技术无法应对数据量急剧增长带来的挑战，分布式技术手段则使计算能力和存储能力获得了充分的增长，解决了基于海量数据的数据挖掘出问题。

1.6 数据科学项目的基本流程

数据科学项目一般从一个问题开始。例如，某区域新楼盘预期的价格是多少？消费者对某商品的态度是正面的还是负面的？某邮件是否是垃圾邮件？

正确理解问题后，要依次进行以下工作。

（1）数据获取。数据的可用性是项目成功必不可少的条件。公开数据可以直接获取。如果数据存在但没有公开，就需要以协商（即使是本单位的数据）或购买的方式合法地获取数据。如果数据不存在，可能就要通过某些设备或程序（如爬虫）对数据进行合法采集并存储起来。有些数据可能包含敏感信息，进入这一阶段之前，要对数据进行脱敏处理，使之能保留原有的意义，又不会造成隐私泄露。

（2）数据预处理。获得的数据可能包含多个数据集。这些数据集的产生时间、保存格式可能是不同的，需要将这些数据集组织成单一的、一致的、高质量的数据集。在实践中，数据（特别是直接采集的数据）还可能包含部分无效数据项，必须应用一些规则防止“脏”数据影响模型。

（3）探索数据特征。在探索数据特征阶段，需要从数据集中提取最适合任务的变量或数据特征。有些特征可能已经存在于数据集中了，有时需要根据多个现有特征

来设计新的数据特征。寻找数据特征的依据是确定哪些变量对问题建模最有用。这一过程既是科学，也是艺术，是一个复杂而重要的过程。

（4）构建和调整模型。构建和调整模型是指选择建模技术，并在数据集中应用该技术。从高层次上区分，有两种类型的建模技术：监督学习和无监督学习。

监督学习需要包含一批样本的训练集，每个样本由特征和目标变量组成。机器学习算法需学习如何将一组特征映射为目标变量的值。回归是常见的有监督学习模型，它的一个典型应用是价格预测：用一组房屋变量（如面积、朝向、距市中心的距离）和对应价格作为监督数据（或称训练集），通过学习获得回归模型，可以预测其他房屋的价格。分类是另一种有监督学习模型。垃圾邮件过滤器是一个机器学习程序。通过学习用户标记好的垃圾邮件和非垃圾邮件示例（训练集），它可以学会自动标记垃圾邮件。

无监督学习建模的目的是识别数据中的模式（规律），不需要任何有标记的训练集。聚类、异常检测属于无监督学习。

模型构建后，需要进行检验。如果没有达到标准，就需要返回上一个步骤去获取新的或不同数据，或者使用不同的特征，来构建新的模型。

1.7 数据安全和大数据伦理

1.7.1 数据安全

数据安全指保护信息系统或数据资源免受各类干扰、破坏和非法访问。数据安全问题是人类社会在信息化发展过程中无法回避的问题。大数据蕴含着巨大的价值，更易成为被攻击的重点目标。近年来，数据泄露等安全问题事件频发。雅虎（Yahoo）、推特（Twitter）、脸书（Facebook）均发生过数据泄露事件。在国内，2020 年青岛胶州某医院出现了个人信息泄露事件，2021 年北京智借网络科技有限公司非法出售用户个人信息，2022 年 6 月美国国家安全局窃取了西北工业大学远程业务操作记录等关键敏感数据。这一系列的安全事件使数以亿计的用户成为受害者，社会影响十分恶劣，不仅给个人和企业带来了威胁，还严重危害了社会安定和国家安全。

数据安全包括传统数据安全和大数据安全两个方面。

1. 传统数据安全

传统上，数据可能受到四方面的威胁。

（1）计算机病毒。计算机病毒会影响软件、硬件的正常工作，破坏数据或窃取数据。

（2）黑客攻击。黑客通过网络，利用目标计算机的漏洞控制目标计算机，窃取、破坏或篡改数据。

（3）介质损坏。数据通过网络传输并存放在某种形式的存储设备上。自然灾害

可能会造成网络中断、设备损坏或灭失；停电可能会造成传输中或存储设备上的信息损坏或丢失。

（4）人为因素。密码保管不善会造成数据的泄露；操作失误可能会造成文件误删、存储设备格式化、设备遗失。

2. 大数据安全

大数据具有共享性，可以交易，其动态利用已逐渐走向常态化和多元化。实现数据价值的渠道往往依赖于大量多样性数据的汇聚、流动、处理和分析活动，而这些活动所涉及的治理主体更加多元，利益诉求更加多样，数据安全概念的内涵和外延均在不断扩充、延展，大数据安全也表现出了新的特征。

（1）大数据成为网络攻击的首要目标。大数据体量大，存放集中，攻击成功后回报较高。

（2）大数据本身推高了数据泄露的风险。各种细节数据不断在某一平台上持续聚集，使攻击者非法获取数据后更容易解读数据，而这些被非法获取的数据可以用来对其他平台进行“撞库”攻击。如图 1.1 所示，攻击者利用 A 平台获取的用户和密码（部分用户喜欢在不同平台使用相同的密码），通过“撞运气”的方式尝试破解平台 B 的相关信息。一个安全保障能力较弱的平台发生数据泄露后，往往会成为攻击其他平台的资源。攻击者可能通过收集用户上网（如社交网络、邮件、微博、快递数据）的痕迹，获取攻击目标的相关信息，提高攻击成功率。

图 1.1 “撞库”攻击

（3）大数据安全关乎国家安全。随着数据价值的不断提高，数据资源成为国家核心战略资产和社会财富。一个国家拥有数据的规模和应用能力将成为综合国力的重要组成部分。数据安全已成为维护国家主权和核心利益的基础。另外，自媒体（如微博、微信和抖音等）使更多的互联网用户获得了独立表达自己观点的能力，但是自媒体的发展良莠不齐。一些自媒体为了单击率，不断突破道德底线，发布虚假消息，误导受众，冲击主流媒体，成为影响国家意识形态安全的因素。

大数据的产生、收集、存储、使用、传输、共享、发布和销毁等阶段都面临着新的安全威胁和挑战。数据加密、用户访问权限控制、数据隔离、数字签名等技术和严格执行相关法律法规是提高数据安全的主要手段。

1.7.2 大数据伦理

伦理（ethics）是指“规则和道理”，其中“规则”（伦）和“道理”（理）是同一概念中的两个方面。美国《韦氏大辞典》指出，伦理学是一门探讨什么是好、什么是坏以及道德责任义务的学科。数据伦理（data ethics）是指在数据生产、治理、使用和共享过程中个人和机构需要遵守的社会道德和科学规范，是数据从业人员和机构应该遵循的职业道德准则。

大数据是一种资源，但可能会被用于技术性垄断，从而获取不当的优势。例如，“大数据杀熟”和“困在算法里的美团骑手”一度成为社会舆论的焦点。“大数据杀熟”是指平台（主要是互联网平台）充分利用自身所掌握的大数据技术，对熟人（忠诚的用户）进行不当的价格歧视，从而使大数据技术成为部分经营者追求不当利润的工具，普通消费者很难通过网络对经营者价格歧视的行为进行甄别。“困在算法里的美团骑手”是指外卖骑手的收入被大数据和算法支配。为了“准时送达”，骑手们工作时经常处于高度紧张的状态，甚至违反交通法规，给自己和他人的生命安全带来了极大的威胁。此外，有些企业通过摄像头对客户进行人脸识别，以便在议价方面取得优势。

大数据的核心是预测，它为人类的生活创造了前所未有的可量化维度，已经成为新发明和新服务的源泉。但是，这种改变一旦缺失对伦理道德的坚守，结果会适得其反，使大数据营销变为算法霸权，直接导致公众利益受损，还会引发公众对企业的信任危机。

需要坚持“科技向善”理念，推动科技的发展和创新，尽可能地减少新技术带来的负面影响，充分引领新科技正向价值的发挥。为了更好地为人类、为社会服务，大数据技术应遵循的伦理原则如下。

（1）无害性原则。大数据技术发展应坚持以人为本，服务于人类社会的健康发展和人民生活质量的提高。

（2）权责统一原则。谁搜集谁负责，谁使用谁负责。

（3）尊重自主原则。数据的存储、删除、使用、知情等权利应充分赋予数据产生者。

对于大数据技术带来的伦理问题，最有效的解决之道就是继续推动技术进步。应鼓励以技术进步消除大数据技术的负面效应，从技术层面提高数据安全管理水平。例如，对个人身份信息、敏感信息等采取数据加密升级和认证保护技术；将隐私保护和信息安全纳入技术开发规范，作为技术原则和标准。

2019 年 1 月 1 日，我国正式施行《电子商务法》，其中第十八条规定：“电子商务经营者根据消费者的兴趣爱好、消费习惯等特征向其提供商品或者服务的搜索结果的，应当同时向该消费者提供不针对其个人特征的选项，尊重和平等保护消费者合法权益。”相关法律规定的出台为消除“大数据杀熟”现象、过度收集信息和保障大数据安全做出了努力，但最根本的还是在于企业组织自身的伦理底线。只有将人文情怀赋予理性工具，在算法背后辅以道德正义的支撑，才能真正实现大数据造福大众的目标。

1.8 国家层面的大数据问题

1.8.1 数据主权

互联网服务“无国界”的性质侵蚀了传统主权和领土管辖权的概念，但土地、水、人口、健康、金融和犯罪等方面的数据对一个国家来讲至关重要，各国正在寻找新的适当方法来保障国家数据安全。数据主权是国家主权的重要组成部分，指以符合数据所在国法律、惯例和习俗的方式管理数据，也指国家采取一系列方法控制在本国互联网基础设施中生成或通过本国互联网基础设施生成的数据，并将数据流置于国家管辖范围内。新一代信息技术催生了大批中国高科技企业，一部分企业掌握着海量的用户敏感数据，在跨国发展过程中可能会损害国家数据安全和公共利益。此外，境外公司在我国境内运营过程中也收集了海量数据，有着极大的安全隐患。我国应当将保护数据安全置于国家战略高度，顺应全球数字经济的严格监管趋势，全力捍卫国家数据主权。

美国的数据主权安全保障及其战略建设开始于 20 世纪 80 年代。作为全球最早开始建设数据主权战略的国家，美国至今已出台了 130 余部相关法案，形成了同时涵盖互联网宏观整体规范与微观具体规定的完备数据主权战略体系。近年来，美国接连出台了《国家网络战略》《澄清境外数据的合法使用法案》和《消费者隐私法案》等相关法规、政策，通过双重标准、长臂管辖等手段对他国企业进行打压，谋求继续主导网络空间的国际治理规则。

欧盟也全力推进数据主权战略的构建。2018 年，欧盟出台了《通用数据保护条例》；2020 年，欧盟通过了《欧洲数据治理条例（数据治理法）》提案，并公布了《数字服务法案》《数字市场法案》两部法案的草案；2021 年 3 月，欧盟委员会发布了《2030 数字罗盘：数字十年的欧洲方式》，提出了未来十年，欧洲加快数字化转型的具体目标以及衡量目标完成情况的数字罗盘。欧盟的数字监管模式极大地影响了世界各国以及各大企业的数据监管措施。企业为了进入欧洲市场，需要主动或被动地将其数据保护措施提升至欧盟标准。

我国于 2016 年颁布的《网络安全法》创建了一个广泛的数据保护框架。出于对

国家安全的考虑，法规中包含了对个人信息和重要数据的数据本地化要求。2017 年，中共中央网络安全和信息化委员会办公室（以下简称中央网信办）发布了《个人信息和重要数据出境安全评估办法》，将数据本地化要求扩展到所有网络运营商，而不仅是《网络安全法》中规定的关键信息基础设施运营商，对数据本地化存储的要求更加严格。在《网络安全法》的框架下，我国又陆续制定了《数据安全法》《个人信息保护法》等法律，出台了《互联网用户账号信息管理规定》《数据出境安全评估办法》等一系列规章制度，要求在数据出境时对运营商进行安全评估，以防存在影响国家安全或损害公共利益的风险，确保我国的数据主权保护得到落实。我国拥有自己的国家内联网，互联网流通内容需要经过审查，审查和约谈机制成为我国独特的数据治理措施。在强化数据跨境流动监管的同时，我国政府也在大力投资境内数字基础设施建设，并通过“数字丝绸之路”计划在全球范围内扩大通信基础设施建设，创造了以中国为中心的跨国网络基础设施体系。

2021 年 12 月，鉴于近期新浪微博及其账号屡次出现法律、法规禁止发布或者传输的信息，且情节严重，国家互联网信息办公室负责人约谈了新浪微博的主要负责人、总编辑，依据《网络安全法》《未成年人保护法》等法律法规，责令其立即整改，并严肃处理相关责任人。北京市互联网信息办公室对新浪微博运营主体北京微梦创科网络技术有限公司依法予以共计 300 万元罚款的行政处罚。除此之外，滴滴全球股份有限公司在经营过程中存在严重影响国家安全的数据处理活动并违规收集了大量的客户敏感数据。2022 年 7 月 21 日，国家互联网信息办公室依据《网络安全法》《数据安全法》《个人信息保护法》《行政处罚法》等法律法规，对滴滴全球股份有限公司处以人民币 80.26 亿元的罚款，对滴滴全球股份有限公司董事长兼 CEO 程维、总裁柳青各处人民币 100 万元的罚款。

数据主权也可以泛指组织和个人对自己产生的各种有价值数据资源的占有、使用、解释、自我管理、自我保护，并且不受任何组织、单位和个人侵犯的权利。组织和个人的数据主权必须无条件地服从国家数据主权的需要，国家数据主权是第一位的。

1.8.2 大数据与国家治理

大数据不仅是一场技术革命，一场经济变革，也是一场国家治理的变革。维克托•迈尔•舍恩伯格在其著作《大数据时代》中说：“大数据是人们获得新的认知、创造新的价值的源泉，还是改变市场、组织机构以及政府与公民关系的方法。”

在大数据时代，互联网是政府施政的新平台。“十三五”规划建议指出：“运用大数据技术，提高经济运行信息及时性和准确性。”大数据正有力地推动着国家治理体系和治理能力走向现代化，正日益成为社会管理的驱动力、政府治理的重要依据。目前，大数据正逐渐成为国家战略。2014 年 7 月 23 日，国务院常务会议审议通过《企业信息公示暂行条例（草案）》，推动构建公平竞争市场环境。其中要求建立部门间互联共享信息平台，运用大数据等手段提升监管水平。2014 年 9 月 17 日，国务院常

务会议部署进一步扶持小微企业发展，推动大众创业，万众创新，其中包括加大服务小微企业的信息系统建设，方便企业获得政策信息，运用大数据、云计算等技术提供更有效的服务。2014 年 10 月 29 日，国务院常务会议要求重点推进 6 大领域消费，其中强调加快健康医疗、企业监管等大数据应用。2014 年 11 月 15 日，国务院常务会议提出在疾病防治、灾害预防、社会保障、电子政务等领域开展大数据应用示范。2015 年 1 月 14 日，国务院常务会议部署加快发展服务贸易，以结构优化拓展发展空间，提出要创新模式，利用大数据、物联网等新技术打造服务贸易新型网络平台。2015 年 2 月 6 日，国务院常务会议确定运用互联网和大数据技术，加快建设投资项目在线审批监管平台，横向联通发展改革、城乡规划、国土资源、环境保护等部门，纵向贯通各级政府，推进网上受理、办理、监管“一条龙”服务，做到全透明、可核查，让信息多跑路，群众少跑腿。2015 年 7 月，国务院办公厅印发的《关于运用大数据加强对市场主体服务和监管的若干意见》提出，要提高对市场主体服务水平；加强和改进市场监管；推进政府和社会信息资源开放共享；提高政府运用大数据的能力；积极培育和发展社会化征信服务。

1.8.3 大数据重塑世界新格局

“数据是新的石油，是本世纪最为珍贵的财产。”大数据正在改变各国综合国力，重塑未来国际战略格局。

大数据正在成为经济社会发展新的驱动力。随着云计算、移动互联网等网络新技术的应用、发展与普及，社会信息化进程进入数据时代，海量数据的产生与流转成为常态，将涵盖经济社会发展的各个领域，成为新的重要驱动力。大数据重新定义了各个大国博弈的空间。在大数据时代，世界各国对数据的依赖性快速上升，国家的竞争焦点已经从资本、土地、人口、资源的争夺转向了对大数据的争夺。未来国家层面的竞争力将部分体现为一国拥有数据的规模、活性以及解释、运用的能力，数字主权将成为继边防、海防、空防之后另一个大国博弈的空间。大数据将改变国家的治理架构和模式。在大数据时代，可以通过对海量、动态、高增长、多元化、多样化数据的高速处理，快速获得有价值的信息，提高公共决策能力。

鉴于大数据潜在的巨大影响，很多国家或国际组织都将大数据视作战略资源，并将大数据提升为国家战略。2012 年 3 月，美国政府发布了“大数据研发计划”，并设立了 2 亿美元的启动资金，希望增强海量数据的收集、分析、萃取能力，认为这事关美国的国家安全和未来竞争力。迄今为止，美国在大数据方面实施了三轮政策，开放了 50 多个门类的政府数据，以确保商业创新。同时，欧盟正在力推《数据价值链战略计划》，为 320 万人增加就业机会；日本积极谋划利用大数据改造国家治理体系，对冲经济下行风险；联合国推出了“全球脉动”项目，希望利用“大数据”预测某些地区的失业率或疾病暴发等现象，以提前指导援助项目。截至 2014 年 4 月，全球已有 63 个国家制定了开放政府数据计划，推动政府从“权威治理”向“数据治理”转变。

中国国际经济交流中心副研究员张茉楠撰文指出，中国需要加快形成大数据国家战略，着力规划大数据战略中长期路线图与实施重点、目标、路径，统筹布局，加快大数据发展核心技术的研发，推进大数据开放、共享以及安全方面的相关立法与标准制定，抢占全球科技革命和产业革命战略机遇期，重构国家综合竞争优势。

1.8.4 中国国家大数据战略

2014 年 3 月，大数据首次写入中国政府工作报告；2015 年 8 月，国务院常务会议通过了《促进大数据发展行动纲要》；同年 10 月，党的十八届五中全会正式提出“实施国家大数据战略，推进数据资源开放共享”；2016 年，我国《国民经济和社会发展第十三个五年规划纲要》正式提出“实施国家大数据战略”；2021 年 12 月，工业和信息化部发布的《“十四五”大数据产业发展规划》指出：要充分激发数据要素的价值潜能，打造数字经济发展的新优势，为建设制造强国、网络强国、数字中国提供有力支撑。目前，我国已经将大数据视作战略资源上升为国家战略，以大数据创新驱动中国特色社会主义各项事业的发展。

我国国家大数据战略的十六字方针是“审时度势、精心谋划、超前布局、力争主动”，具有以下特征。

（1）时代性。“审时度势”要求准确把握当前大数据的发展现状和趋势，对大数据的战略发展机遇和面临的挑战要有清醒的认知。

（2）系统性。“精心谋划”要求围绕网络强国、数字中国、智慧社会的建设目标，加强国家大数据战略的科学实施。将大数据战略与创新驱动发展战略、网络强国战略等进行系统设计与统筹推进，是我国大数据战略实施的重要特点。同时，国家大数据战略的系统性，不仅体现在从国家层面对大数据的技术创新、平台建设、制度管理、人才培养、标准制定等进行整体设计与推进，也体现在对各地方、各行业等的大数据发展进行统筹协调与推进。

（3）超前性。“超前布局”就是要在着力解决制约大数据发展的现实问题的基础上，立足于大数据的未来发展，进行前瞻性布局。如通过制定和实施大数据发展规划，明确大数据产业发展的主要目标、任务、计划和政策措施。继 2015 年国务院印发《大数据发展行动纲要》后，2016 年工业与信息化部发布了《大数据产业发展规划（2016—2020）》，2019 年国务院印发了《新一代人工智能发展规划》等，对我国大数据发展如何保持前沿性、先进性等做出了具体部署。

（4）自主性。“力争主动”就是要把握信息化发展进入大数据新阶段的时间窗口，充分发挥中国网络大国的优势，借鉴世界各国的发展经验，走中国特色的大数据战略发展和网络强国之路。与英美等发达国家相比，我国大数据发展存在信息基础设施建设明显滞后、网络安全面临严峻挑战、一些关键核心技术受制于人等问题。因此，要发挥我国的制度优势和市场优势，面向国家重大需求，面向国民经济发展主战场，全面实施促进大数据发展的行动。

大数据是每个人的大数据，是每个企业的大数据，更是整个国家的大数据。随着国家大数据战略的实施，基于大数据的智慧生活、智慧企业、智慧城市、智慧政府、智慧国家必将一一实现。

1.9 云计算

服务器是计算机的一种，它们在网络中为其他客户机（如 PC 机、智能手机、ATM 等终端甚至是火车系统等大型设备）提供计算、存储服务。服务器具有强大的运算能力，能长时间地可靠运行，具备强大的输入 / 输出能力以及更好的扩展性。

最初，每个应用程序（application，为用户直接提供服务的软件）都要建立自己的服务器体系。如图 1.2 所示，应用程序 A 使用了 4 台服务器，应用程序 B 使用了 5 台服务器。这种方式的缺点是两组服务器的计算能力（以下简称算力）不能共享，在某一时段可能一个应用程序因负载过大影响了服务质量，另一个的负载却很小，造成了算力浪费。

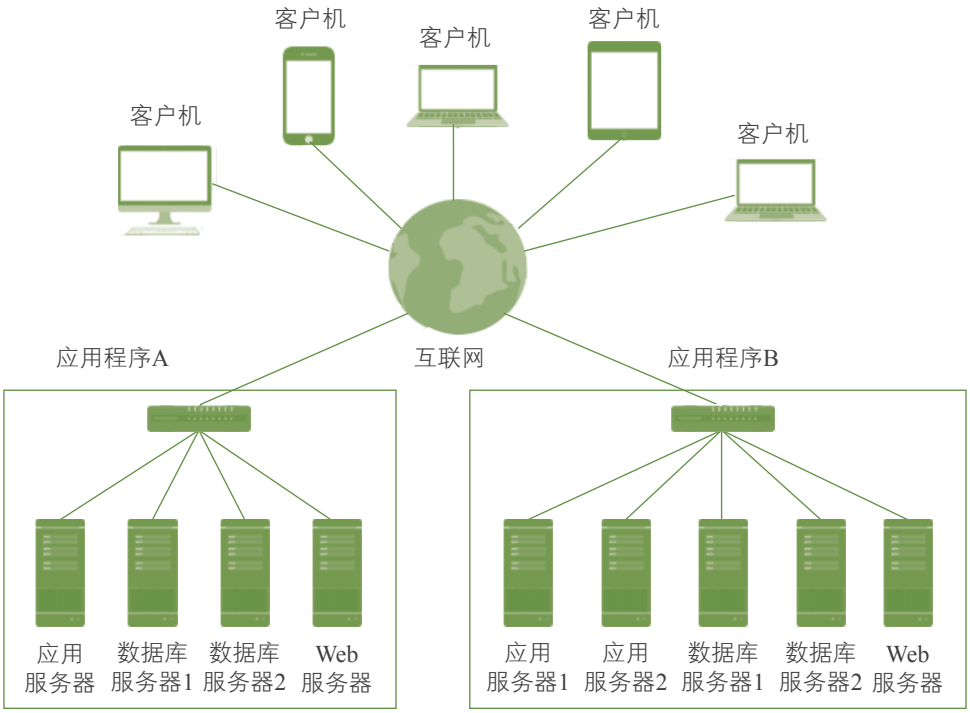

图 1.2　服务器和客户机

鉴于这种情况，云计算应运而生。云计算能提供可伸缩的、廉价的计算能力服务，使用者可以随时获取“云”上的资源，按需求量使用，按使用量付费。算力就像自来水厂提供的水一样（背后可能连接不同的水库），用户可以随时接水，并按用水量付费就可以。在图 1.3 所示的云计算平台和客户机的关系示意图中（服务器被有意虚化），各个应用程序不再独占某一组服务器资源，而是虚拟地使用云计算平台提供的算力。

某个应用程序负载过大时，可以购买更多的算力；而负载下降时，可以减少算力购买量。

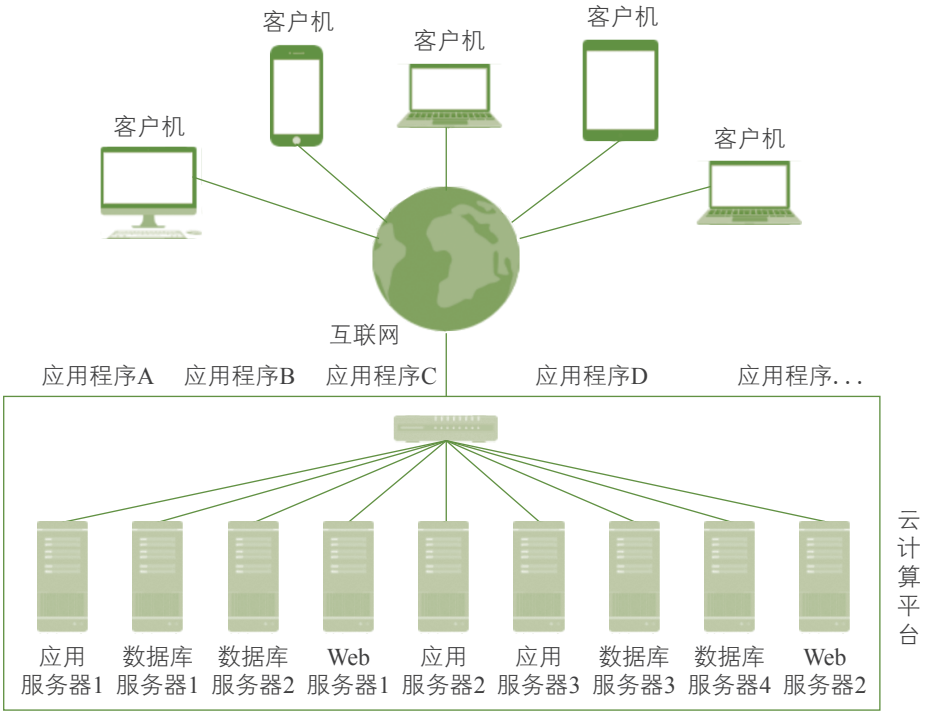

图 1.3　云计算平台和客户机

云计算不是一种全新的网络技术，而是一种全新的网络应用理念。云计算的核心概念是以互联网为中心，在网站上提供快速且安全的计算服务与数据存储，让每一个使用互联网的人都可以使用网络上的庞大计算资源与数据中心。

2006 年 3 月，亚马逊云服务（Amazon Web Services，AWS）发布了 Amazon Simple Storage Service(Amazon S3)，开始以 Web 服务的形式向企业提供 IT 基础设施服务（infrastructure as a service，IaaS），开创了一种崭新的计算资源服务模式。彼时还没有“云计算”这个名称。同年 8 月，Google 首席执行官埃里克 • 施密特在搜索引擎大会上首次提出了“云计算”（cloud computing）的概念，这是云计算发展史上第一次正式地提出这一概念。

关于云计算，美国国家标准与技术研究院（national institute of standards and technology，NIST）给出的定义是：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问。用户进入可配置的计算资源共享池（资源包括网络、服务器、存储、应用软件）后，这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。云计算指特定的计算能力服务模式，也可以指实现该模式的各种技术。如果不特别说明，本书中的“云计算”指这种算力提供模式。

1.9.1 云计算的特征

云计算具有以下特征。

（1）超大规模的算力。云计算平台将存储和运算能力分布在网络所连接的各个节点之中，计算架构由“服务器 + 客户端”向“云服务平台 + 客户端”演进。企业内部的云平台有数百台服务器协同工作，而经营性的云平台更是拥有数十万甚至数百万台服务器，对外提供超大规模的算力。

（2）高可靠性。云计算平台由大量计算机组成集群向用户提供数据处理服务，利用多种硬件和软件冗余机制，使用数据多副本容错、计算节点同构可互换等措施，部分软件或硬件出现故障时，仍然可以持续地对外的服务。冗余的 IT 资源还被部署在不同的物理位置，即使某一地域出现灾难性事件，也不会影响服务。

（3）灵活性。云计算平台服务的实现机制对用户透明，用户无须了解云计算的具体机制，就可以获得需要的服务。用户可以在任何位置，利用具有互联网访问功能的设备，如 PC 或者智能手机，通过互联网访问他们所需的信息，获得他们所需的服务。

（4）按需使用。云计算的基础设施通常是以算力的形式提供服务，这使用户不需要为了一次性或非经常性的计算任务购买昂贵的设备，而是“租用”计算资源。当用户需要更多算力时，就会购买更多的服务；业务量小时，就可以购买较少资源。这个特征也被称为服务的“弹性”。

（5）多租户模式。现在大部分的软件和硬件都对虚拟化有一定支持，各种资源、软件、硬件都虚拟化放在云计算平台中统一管理。使用平台服务的租户之间是隔离的。即使他们在同时使用某一相同的 IT 资源，也不会相互影响。

1.9.2 云计算的典型服务模式

云计算包括三种典型的服务模式：基础设施即服务、平台即服务和软件即服务。

（1）基础设施即服务。最早的云计算服务是基础设施服务（infrastructure as a service，IaaS）。平台提供主机、存储、网络和安全等几个重要的基础云服务，允许用户灵活组合，并实现了弹性计费，即用户可以按时间段租用云主机。租用主机的客户不需要自行购置服务器，但需要安装操作系统、数据库、应用服务器，还需要开发应用程序，部署运维，才能实现云端应用。

（2）平台即服务。平台即服务（platform as a service，PaaS）模式为客户提供了一个完整的云平台（如硬件、软件和基础架构），用于开发、运行和管理应用程序，而无须考虑在本地构建和维护该平台通常会带来的成本、复杂性和不灵活性。PaaS 提供商将服务器、网络、存储等基础设施服务及操作系统软件、数据库、应用服务器、开发工具等一切工具都托管在其数据中心上。通常，客户可以支付固定费用，为指定数量的用户提供指定数量的资源；也可以选择“按使用量付费”定价模式，仅为他们使用的资源付费。租用主机的客户不需要自行购置服务器，也无须安装操作系统、数

据库、应用服务器，只需要开发应用程序，部署运维，就能实现云端应用。

（3）软件即服务。软件即服务（software as a service，SaaS）是一种基于互联网提供软件程序功能（而不再提供软件产品）的应用模式。SaaS 建立在 IaaS 及 PaaS 的基础之上，是云服务中最上层、直面用户的一层。SaaS 模式改变了传统软件服务的提供方式，减少了本地部署所需的大量前期投入，进一步突出了软件的服务属性，也使软件进一步回归服务本质。

长期以来，字处理软件都是以软件产品的方式存在的，客户需要下载、安装、运行软件，才能使用。近几年来，采用 SaaS 模式的在线文档迅速发展起来，客户无须软件产品即可获得服务。此外，一些传统的软件也提供了 SaaS 模式，体现了社会对 SaaS 模式的认可。可以预见，SaaS 模式将会成为软件的主流模式。

1.9.3 云计算服务部署的环境

按云计算服务部署的环境来划分，云计算有公有云、私有云和混合云三种类型。

（1）公有云。公有云的核心属性是共享资源服务，云提供者创建并维护公有云的 IT 资源，对外部用户提供服务。如华为云、阿里云、腾讯云和 Amazon AWS 都为组织外用户提供服务。

（2）私有云。私有云只为组织内部的用户提供服务。在私有云模式下，组织把云计算技术当作一种手段，可以集中访问不同部分、不同部门的 IT 资源。由于私有云可控性更强，一些大型企业会出于安全考虑自建云环境，只为企业内部提供服务。尽管不对组织外的用户提供服务，但它仍然采用了云计算技术，所以它仍然是云服务。

（3）混合云。混合云综合了公有云和私有云的特点。一些企业出于安全考虑，把处理敏感数据的云服务部署到私有云上，一般服务部署到公有云上，把公有云和私有云进行混合搭配使用。

1.9.4 云计算和大数据的关系

大数据必然无法用单台的计算机进行处理，必须采用分布式计算架构，必须依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术。云计算降低了计算资源的成本和技术壁垒，为大数据技术发展奠定了坚实的算力基础。大数据和云计算之间的关系就像容器和水的关系，云计算就像一个容器，而大数据则正是存放在这个容器中的水。

1.10 物联网

1. 物联网概述

物联网的概念最早出现于比尔•盖茨 1995 年出版的《未来之路》一书。在《未来之路》中，比尔 • 盖茨已经提及物联网概念，只是当时受限于传感设备、无线网络及硬件的发展，并未引起世人的重视。1998 年，美国麻省理工学院创造性地提出了当时被称作 EPC（electronic product code，产品电子代码）系统的“物联网”构想。1999 年，美

国自动识别中心（AutoID Center，由麻省理工学院创建）正式提出了“物联网”的概念，主要建立在物品编码、射频识别技术和互联网的基础上。

物联网（internet of things，IoT）是新一代信息技术的重要组成部分，具有广泛的用途，和云计算、大数据有着紧密的联系。

物联网是物物相连的互联网，是互联网的延伸。它利用局域网或互联网等通信技术把传感器、智能设备、计算机、人员和物料联系在一起，形成人与物相连、物与物相连，实现数据的自动采集、自动传输、自动处理。传感器和智能设备的高速发展，使人们能够以低成本、高效率的方式实现对机器数据的大规模采集。

从技术架构上看，物联网可以分为四个层次，即感知层、网络层、处理层和应用层。假设某楼宇有门禁卡和人脸识别两种身份验证方式，则验证工作流程会依次涉及这四层。

（1）感知层。感知层负责获得数据，门禁卡和人脸识别摄像头可以分别获取卡号或人脸数据。

（2）网络层。网络层负责传输数据，卡号或人脸图像通过网络（无线或有线）进行传输。

（3）处理层。处理层负责处理数据，程序比对预先在数据库中保存的卡号或人脸信息，并返回“开门”或“无权限”信号。该信号通过网络层传输到门的控制设备后，设备执行开门或拒绝开门的动作。此外，处理层还有安全管理、网络管理等职能。

（4）应用层。应用层负责和用户交互，可以让管理员上传图片、输入卡号、查看设备状态等。

2. 物联网的应用

图 1.4 是智慧公交的一个界面，用户可以查看公交车的实时位置及交通状况，还可以付款。公交车都安装有定位系统（北斗或 GPS）和 4G/5G 通信设备，行驶过程中会将位置数据通过通信设备发送到公交指挥中心，指挥中心随时更新数据库中的数据，用户通过手机或其他联网的设备即可访问公交车的实时数据。

图 1.4　智慧公交界面

物联网的应用领域涉及众多行业，有效地推动了这些行业的智能化发展，使有限的资源可以更合理的方式进行分配，从而提高了行业效率和效益，大大改善了人们的生活质量。

（1）基础设施领域。目前，交通拥堵已成为城市的一大问题。对此，可通过交通物联网对道路交通状况进行实时监控，并将信息及时传递给驾驶人，让驾驶人及时做出出行调整，

以缓解交通压力；在高速路口设置道路 ETC（自动收费系统），免去进出口取卡、还卡的时间，可提升车辆的通行效率；在公交车上安装定位系统，能及时了解公交车的行驶路线及到站时间，乘客可以根据出行计划选择乘车路线。另外，不少城市推出了智慧路边停车管理系统，基于云计算平台，结合物联网技术与移动支付技术，共享车位资源，提高了车位利用率，方便了用户。

（2）公共安全领域。近年来，全球气候异常情况频发，灾害的突发性和危害性进一步加大。对此，可使用物联网实时监测环境的安全情况，实现实时预警，使人们可提前预防，及时采取应对措施，降低灾害对人类生命财产的威胁。例如，将通过特殊处理的感应装置置于深海，可分析水下相关情况，实现对海洋污染的防治、海底资源的探测等，甚至对海啸也可以提供更加可靠的预警；利用物联网技术可以智能感知大气、土壤、森林、水资源等方面的指标数据，改善人类的生活环境。

（3）智能家居领域。随着宽带业务的普及，即使家中无人，也可利用手机等产品客户端远程操作智能空调调节室温，实现智能灯泡的开关，调控灯泡的亮度和颜色等；插座内置的 Wifi 可实现遥控插座的定时通断电流，甚至可以监测设备用电情况，生成用电图表，安排资源使用及开支预算等。另外，智能摄像头、窗户传感器、智能门铃、烟雾探测器、智能报警器等都是家庭可安装的物联网监控设备。

（4）企业和事业单位。在畜牧领域，可以将牲畜的活动信息、生理检测信息发送至云平台，可实现场舍温度控制、设备、消杀管理、系统管理及统计报表生成等；在环保领域，可以应用物联网技术进行环境质量监测以及可视化呈现；在工业制造领域，通过物联网技术在产线上添加多个传感器，获取生产线上的实时合格率，结合对应产出的销售数据，从而得到该产线的实时效能、实时毛利率等运营数据，让工厂效能最大化。此外，物联网还在智慧校园、数字政府等事业和机关单位的智能管理中得到了深入的应用。

（5）国防军事领域。大到卫星、导弹、飞机、潜艇等装备系统，小到单兵作战装备，物联网技术的嵌入有效提升了军事智能化、信息化、精准化，极大提升了军事战斗力，是未来军事变革的关键。

物联网中每时每刻都在产生、传输海量数据，是大数据的主要来源之一。没有物联网的飞速发展，就不会带来数据产生方式的变革（由人工产生阶段转向自动产生阶段），人类社会也不会这么快进入大数据时代。

1.11 数字经济

数字经济是以数据资源为关键要素，以现代信息网络为主要载体，以信息通信技术融合应用、全要素数字化转型为重要推动力，促进公平与效率更加统一的新经济形态，最早出现在 20 世纪 90 年代。伴随着互联网、大数据、5G、人工智能等为代表的新一

代数字技术的不断革新，数字经济得到了迅速发展，成为世界经济增长的重要驱动力。

1996 年，美国学者唐·泰普斯科特（Don Tapscott）出版的《数字经济：网络智能时代的前景与风险》描述了互联网将如何改变世界各类事务的运行模式并引发若干新的经济形式和活动，第一次提出了“数字经济”这一概念。2002 年，美国学者金范秀（Beomsoo Kim）将数字经济定义为一种特殊的经济形态，其本质为“商品和服务以信息化形式进行交易”。当时的信息技术对经济的影响尚未具备颠覆性，只是提质增效的一种手段，数字经济并没有引起全社会的共同关注。

大数据时代为数字经济赋予了新的含义。2016 年 9 月，二十国集团领导人杭州峰会通过的《二十国集团数字经济发展与合作倡议》中指出，数字经济是指以使用数字化的知识和信息作为关键生产要素，以现代信息网络作为重要载体，以信息通信技术的有效使用作为效率提升和经济结构优化的重要推动力的一系列经济活动。

通常把数字经济分为数字产业化和产业数字化两方面。数字产业化指信息技术产业的发展，包括电子信息制造业、软件和信息服务业、信息通信业等数字相关产业；产业数字化指以新一代信息技术为支撑，对传统产业及其产业链上下游进行全要素的数字化改造，通过与信息技术的深度融合，实现赋值、赋能。从外延看，经济发展离不开社会发展，社会的数字化无疑是数字经济发展的土壤，数字政府、数字社会、数字治理体系建设等构成了数字经济发展的环境；同时，数字基础设施建设以及传统物理基础设施的数字化奠定了数字经济发展的基础。

数字经济呈现出三个重要特征：

（1）信息化引领。信息技术深度渗入各个行业，促成其数字化并积累大量数据资源，进而通过网络平台实现共享和汇聚，通过挖掘数据、萃取知识和凝练智慧，又使行业变得更加智能。

（2）开放化融合。通过数据的开放、共享与流动，促进组织内各部门间、价值链上各企业间、跨价值链跨行业的不同组织间开展大规模协作和跨界融合，实现价值链的优化与重组。

（3）泛在化普惠。无处不在的信息基础设施、按需服务的云模式和各种商贸、金融等服务平台降低了参与经济活动的门槛，使数字经济出现“人人参与、共建共享”的普惠格局。

数字经济发展速度之快、辐射范围之广、影响程度之深前所未有，正在成为重组全球要素资源、重塑全球经济结构、改变全球竞争格局的关键力量。

1.11.1 大数据与数字经济

1. 大数据开启信息化新阶段，催生数字经济

大数据作为一种概念和思潮在计算领域开始，之后逐渐延伸到科学和商业领域。近 10 年来，大数据相关技术、产品、应用和标准快速发展，逐渐形成了覆盖数据基

础设施、数据分析、数据应用、数据资源、开源平台与工具等板块的大数据产业格局，经历了从基础技术和基础设施、分析方法与技术、行业领域应用、大数据治理到数据生态体系的变迁。

大数据提供了一种人类认识复杂系统的新思维和新手段。理论上来讲，在足够小的时间和空间尺度上对现实世界数字化，可以构造现实世界的一个数字虚拟映像，该映像承载了现实世界的运行规律。在给定充足计算能力和高效数据分析方法的前提下，对这个数字映像的深度分析，将有可能理解和发现现实复杂系统的运行行为、状态和规律。大数据为人类提供了全新的思维方式和探知客观规律、改造自然及社会的新手段，这也是其引发经济社会变革的根本原因之一。

2. 大数据是数字经济的关键生产要素

随着信息通信技术的广泛运用以及新模式、新业态的不断涌现，人类的社会生产生活方式正在发生深刻的变革。数字经济作为一种全新的社会经济形态，正逐渐成为全球经济增长日益重要的驱动力。历史证明，每一次人类社会重大的经济形态变革，必然产生新的生产要素，形成先进生产力。如同农业时代以土地和劳动力、工业时代以资本为新的生产要素一样，数字经济作为继农业经济、工业经济之后的一种新兴经济社会发展形态，也将产生新的生产要素。

数字经济与农业经济、工业经济不同，它是以新一代信息技术为基础，以海量数据的互联和应用为核心，将数据资源融入产业创新和升级各个环节的新经济形态。一方面信息技术与经济社会交汇融合，特别是物联网产业的发展引发数据迅猛增长，大数据已成为社会基础性战略资源，蕴藏着巨大的潜力和能量；另一方面数据资源与产业的交汇融合促使社会生产力发生新的飞跃，大数据成为驱动整个社会运行和经济发展的新兴生产要素，在生产过程中与劳动力、土地、资本等其他生产要素协同创造社会价值。相比其他生产要素，数据资源具有的可复制、可共享、可无限增长和供给的禀赋，打破了自然资源有限供给对增长的制约，为持续增长和永续发展提供了基础与可能，成为数字经济发展的关键生产要素和重要资源。

3. 大数据是发挥数据价值的使能因素

市场经济要求生产要素商品化，以商品形式在市场上通过交易实现流动和配置，从而形成各种生产要素市场。大数据作为数字经济的关键生产要素，构建数据要素市场是发挥市场在资源配置中的决定性作用的必要条件，是发展数字经济的必然要求。2015 年发布的《促进大数据发展行动纲要》明确提出“要引导培育大数据交易市场，开展面向应用的数据交易市场试点，探索开展大数据衍生产品交易，鼓励产业链各环节的市场主体进行数据交换和交易。”大数据发展将重点推进数据流通标准和数据交易体系建设，促进数据交易、共享、转移等环节的规范有序，为构建数据要素市场、实现数据要素的市场化和自由流动提供了可能，成为优化数据要素配置、发挥数据要素价值的关键影响因素。

大数据资源更深层次的处理和应用仍然需要使用大数据，通过大数据分析将数据转换为可用信息，是数据作为关键生产要素实现价值创造的路径演进和必然结果。从构建要素市场、实现生产要素市场化流动到数据的清洗分析、数据要素的市场价值提升和自身价值创造，无不需要大数据作为支撑，大数据已成为发挥数据价值的使能因素。

4. 大数据是驱动数字经济创新发展的动能

推动大数据在社会经济各领域的广泛应用，加快传统产业数字化、智能化，催生数据驱动的新兴业态，能够为我国经济转型发展提供新动力。大数据是驱动数字经济创新发展的重要抓手和核心动能。

大数据驱动传统产业向数字化和智能化方向转型升级，是数字经济推动效率提升和经济结构优化的重要抓手。大数据加速渗透和应用到社会经济的各个领域，通过与传统产业进行深度融合，提升传统产业的生产效率和自主创新能力，深刻变革传统产业的生产方式和管理、营销模式，驱动传统产业实现数字化转型。电信、金融、交通等服务行业利用大数据探索客户细分、风险防控、信用评价等应用，加快业务创新和产业升级步伐。工业大数据贯穿于工业的设计、工艺、生产、管理、服务等各个环节，使工业系统具备描述、诊断、预测、决策、控制等智能化功能，推动工业走向智能化。利用大数据为作物栽培、气候分析等农业生产决策提供有力依据，提高农业生产效率，推动农业向数据驱动的智慧生产方式转型。大数据为传统产业的创新转型、优化升级提供了重要支撑，引领和驱动传统产业实现数字化转型，推动传统经济模式向形态更高级、分工更优化、结构更合理的数字经济模式演进。

大数据推动不同产业之间的融合创新，催生新业态与新模式不断涌现，是数字经济创新驱动能力的重要体现。首先，大数据产业自身催生出如数据交易、数据租赁服务、分析预测服务、决策外包服务等新兴产业业态，同时推动可穿戴设备等智能终端产品的升级，促进电子信息产业提速发展。其次，大数据与行业应用领域深度融合和创新，使传统产业在经营模式、盈利模式和服务模式等方面发生变革，大数据应用已经从通用转向行业应用时代。

基于大数据的创新创业日趋活跃，大数据技术、产业与服务成为社会资本投入的热点。大数据的共享开放成为促进“大众创业、万众创新”的新动力。由技术创新和技术驱动的经济创新是数字经济实现经济包容性增长和发展的关键驱动力。随着大数据技术被广泛接受和应用，诞生出的各种新产业、新消费、新组织形态以及随之而来的创业创新浪潮、产业转型升级、就业结构改善、经济提质增效，正是数字经济的内在要求及创新驱动能力的重要体现。

大数据是数字经济的核心内容和重要驱动力，数字经济是大数据价值的全方位体现。展望未来，要勇于突破、深入探索，应用大数据创造更多新价值，加快产业提质增效，培育壮大经济发展新动能，做大做强数字经济，拓展经济发展新空间，推动经济可持续发展和转型升级。

1.11.2 进一步推动我国数字经济发展

我国数字经济发展迅猛，新产品、新业态、新模式层出不穷，成为驱动中国经济发展的新引擎。习近平总书记指出，“信息化为中华民族带来了千载难逢的机遇”；“发展数字经济意义重大，是把握新一轮科技革命和产业变革新机遇的战略选择”。要牢牢把握机遇，积极应对挑战，克服发展障碍，推进数字经济繁荣发展。为此，应从以下几方面进行努力。

1. 加快数据要素市场培育，激活数据要素潜能

我国已经正式实施《数据安全法》和《个人信息保护法》，为数字经济发展提供了底线保障。为加快数据要素市场培育，还需进一步研究推进数据确权、交易流通、跨境流动等相关制度法规制的修订工作，厘清政府、行业、组织等多方在数据要素市场中的权责边界，同时加强理论研究和技术研发，为数据确权、互操作、共享流通、数据安全、隐私保护等提供有效技术支撑。当前，打破信息孤岛、盘活数据存量是一项紧迫任务，特别是在政务数据领域，应逻辑互联先行，物理集中跟进，完善数据注册、分类分级、质量保障等管理制度和标准规范，在一定层级上构建物理分散、逻辑统一、管控可信、标准一致的政务数据资源共享交换体系，在不改变现有信息系统与数据资源所有权及管理格局的前提下，明晰责权利，确保数据资源高效共享和利用。鼓励在有条件的地区开展数据要素化的探索性实践，鼓励数据运营加工的新业态尝试，以市场化方式推进数据要素市场培育。

2. 推进各行各业的数字化转型

习近平总书记指出，数字经济具有高创新性、强渗透性、广覆盖性，不仅是新的经济增长点，而且是改造提升传统产业的支点，可以成为构建现代化经济体系的重要引擎。当前，信息技术已从助力其他行业提质增效的“工具、助手”角色，转向“主导、引领”角色，深入渗透各个行业，对其生产模式、组织方式和产业形态造成颠覆性影响。然而，面对数字化转型的要求，一些企业却存在“不想、不敢、不会”的“三不”现象。“不想”是囿于传统观念和路径依赖，对新技术应用持抵触情绪；“不敢”是面对转型可能带来的阵痛期和风险，不敢率先探索，就地观望、踌躇徘徊；“不会”则是缺少方法、技术和人才以及成功经验和路径。转型发展必然会面临观念、制度、管理、技术、人才等方面的挑战，其中观念上的转变最为核心和关键，而人才供给则是根本保障。数字化转型并非通过信息技术和工具的简单叠加便可完成，需深度理解“数字化转型、网络化重构、智能化提升”的内涵并系统规划，需要从国家、高校科研院所、企业、社会等多层面打造适应数字化转型需求的数字化人才培养体系，为未来数十年的转型发展储备合格人才。

3. 完善数字治理体系

习近平总书记指出，要完善数字经济治理体系，健全法律法规和政策制度，完善

体制机制，提高我国数字经济治理体系和治理能力的现代化水平。传统的治理体系、机制与规则难以适应数字化发展所带来的变革，无法有效解决数字平台崛起所带来的市场垄断、税收侵蚀、安全隐私、伦理道德等问题，需尽快构建数字治理体系。这其中，数字经济治理无疑是核心内容之一。数字治理体系的构建是一个长期迭代过程，其中，数据治理体系的构建要先行。数据治理体系建设涉及国家、行业和组织三个层次，包含数据的资产地位确立、管理体制机制、共享与开放、安全与隐私保护等内容，需要从制度法规、标准规范、应用实践和支撑技术等方面多管齐下，提供支撑。当前国际数字治理体系尚处于探索期，既有全球性多边机制，也有区域性或双边机制，更有私营平台企业的事实性规则。由于各国数字治理的关注重点不同、发展程度有差异，未来全球数字治理体系将呈现面向关注点差异的、多元化层次化的、多机制共存的格局。

4. 构建“开放创新”“互惠互利”的全球合作伙伴关系

开放创新的本质是从封闭的“机械化思维”到开放的“计算思维”“互联网思维”和“大数据思维”，从“零和博弈”到“协作共赢”。彻底改变了全球软件产业格局的开源软件，是技术领域开放创新最早、最成功的实践。面对数字经济领域的新形势、新任务，需建立互惠互利的合作方式，积极推动国际合作并筹划布局跨国数据共享机制与合理的数据跨境流动机制，与其他国家一起分享数字经济的红利，使我国获得更多发展机遇和更大发展空间。

5. 开展大数据核心关键技术的研发与应用

习近平总书记强调，要加强关键核心技术攻关，牵住数字关键核心技术自主创新这个“牛鼻子”，把发展数字经济自主权牢牢掌握在自己手中。当前，我国仍面临着大数据核心技术受制于人的困境，高端芯片、操作系统、工业设计软件等均是我国被“卡脖子”的短板，需要坚定不移地走自主创新之路，加大力度解决自主可控问题。同时，应针对“人机物”三元融合的万物智能互联时代带来的新需求，把握前沿发展趋势，研发引领性技术，锻造我国的技术长板。核心关键技术大都具有投入高、耗时长、难度大的特点，必须形成科学的管理体制机制，按照创新发展规律、科技管理规律、人才成长规律办事，加强创新资源统筹，优化资源配置，努力取得实质性突破，保障数字经济安全发展。

本章小结

在移动互联网和物联网深入应用的背景下，用户和设备成为产生数据的主流方式。海量数据突破了传统技术的处理能力，并使人类社会进入了大数据时代。大数据带来了数据获取方式、存储方式和处理方式的技术革命，也带来了“总体而非抽样”“效率而非精确”和“相关而非因果”的思维方式的变革。基于大量数据，人们可以通过数据挖掘获得知识，而不局限传统的数学定理和公式。机器学习是根据大量数据获得

规律、构建模型的过程。大数据在带来巨大价值的同时，也使安全问题更加突出，并带来了数据伦理问题，而技术进步和法律健全是解决安全和伦理问题、让大数据造福全人类的有效方式。互联网服务“无国界”的性质侵蚀了传统主权和领土管辖权的概念，但与国家安全有关的数据对一个国家的重要性不亚于传统的战略资源。我国应当将保护数据安全置于国家战略高度，顺应全球数字经济的严格监管趋势，全力捍卫国家数据主权。大数据正在改变国家治理的方式，重塑世界格局，我国已经将大数据视作战略资源上升为国家战略，以大数据创新驱动中国特色社会主义各项事业的发展。

云计算降低了计算资源的成本和技术壁垒，解决了大数据处理因为存储计算资源不足所带来的问题。物联网中每时每刻都在产生、传输着海量数据，是大数据的主要来源之一。没有物联网的飞速发展，就不会带来数据产生方式的变革。

随着新一代数字技术的不断革新，数字经济得到了迅速发展，成为世界经济增长的重要驱动力。大数据时代赋予了数字经济新的含义。数字经济正在成为重组全球要素资源、重塑全球经济结构、改变全球竞争格局的关键力量。

建立数字中国　发展数字经济

党的“二十大”报告中指出，“建设现代化产业体系。坚持把发展经济的着力点放在实体经济上，推进新型工业化，加快建设制造强国、质量强国、航天强国、交通强国、网络强国、数字中国。实施产业基础再造工程和重大技术装备攻关工程，支持专精特新企业发展，推动制造业高端化、智能化、绿色化发展。”实施大数据战略，建立数字中国，在发展实体经济、推进新型工业化、加快建设制造强国中具有基础性的作用。

报告中还提出，“加快发展物联网，建设高效顺畅的流通体系，降低物流成本。”物流是经济增长“主动脉”和“微循环”的重要力量，也是促进国内国际双循环的重要推动力。新冠疫情发生以来，物流行业在抗疫保供、保通保畅、复工复产等方面作用凸显，有效保障了产业链供应链稳定，为经济发展和人民生活提供了重要保障。物联网是大数据的重要源泉，也是现代企业的神经网络。当前全球经济复苏乏力，但随着我国新冠疫情政策的不断优化调整，我国经济逐步回归常态运行。基于物联网、大数据和人工智能的智慧物流需要引领行业创新发展、助力市场保供和产业链的稳定、实现降耗节能、低碳绿色转型，保障“双碳”目标达成。

关于数字经济，报告中提出，“加快发展数字经济，促进数字经济和实体经济深度融合，打造具有国际竞争力的数字产业集群。优化基础设施布局、结构、功能和系统集成，构建现代化基础设施体系。”数字经济赋能传统产业，提供高质量发展重要推动力。数字经济的普惠本质，还是“以人民为中心的发展思想”的落脚点，“发展成果由人民共享”的保证。

随着大数据、云计算、物联网、区块链等前沿信息技术的快速发展，数字技术和数字经济日益成为新一轮国际竞争的重点领域。在全面建设社会主义现代化国家的新征程上，我们需要加快发展数字经济，助推中国经济高质量发展。

习题

1. 请简述大数据的概念和特征。
2. 什么是数据密集型研究范式？
3. 大数据对思维方式产生了什么影响？
4. 请举例说明大数据的安全问题和伦理问题。
5. 请简述数据主权及我国的大数据战略。
6. 请简述云计算、物联网及它们与大数据的关系。
7. 请简述数字经济和大数据对数字经济的意义。